DOCUMENT RESUME

ED 456 129 TM 033 186

AUTHOR Miller-Whitehead, Marie
TITLE Using Testlets To Identify Cognitive Domains Measured by a Kindergarten Diagnostic Assessment.
PUB DATE 2001-06-03
NOTE 25p.
PUB TYPE Reports - Research (143)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Diagnostic Tests; Early Childhood Education; Kindergarten; *Preschool Children; *School Readiness; *Test Use
IDENTIFIERS *Cognitive Domain; *Testlets

ABSTRACT

This study examined a 20-item diagnostic assessment designed to assess readiness of preschool children for kindergarten. Examination of student scores on the individual items of the assessment led to the proposal of the extraction of testlets composed of groups of related items to use information about students' scores better. This approach results in an assessment that essentially is composed of several one-dimensional testlets or blocks that can be scored as scales for achievement or ability on one domain, such as verbal skills. Several simple testlet models, created by principal component analysis and tested by confirmatory factor analysis, are presented, interpreted for the underlying construct, and evaluated. (Contains 5 tables and 40 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 456 129



# Using Testlets to Identify Cognitive Domains Measured by a Kindergarten Diagnostic Assessment

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY HAS BEEN GRANTED BY

M. Miller-Whitehead

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Marie Miller-Whitehead
Tennessee Valley Educators for Excellence
TVEE.ORG
June 3, 2001

TM033186

Marie Miller-Whitehead, Ph.D.
Director
Tennessee Valley Educators for Excellence, Inc.
PO Box 2882
Muscle Shoals, AL 35662

TVEE.ORG
PO Box 682431
Franklin, TN 37068



BEST COPY AVAILABLE

# Using Testlets to Identify Cognitive Domains Measured by a Kindergarten Diagnostic Assessment

Marie Miller-Whitehead

Abstract

This paper examines a 20 item diagnostic assessment designed to assess readiness of preschool children for kindergarten. Examination of student scores on the individual items of the assessment led to the proposal of the extraction of testlets composed of groups of related items to better utilize information about students' scores. This approach results in an assessment that is essentially composed of several unidimensional testlets or blocks that can be scored as scales for achievement or ability on one domain, such as verbal skills. Several simple testlet models, created by Principle Component Analysis and tested by Confirmatory Factor Analysis, are presented, interpreted for underlying construct, and evaluated.

# Using Testlets to Identify Cognitive Domains Measured by a Kindergarten Diagnostic Assessment

## Review of Related Literature

Although most teachers are aware of principles of test design, in the classroom setting it is not always possible or practical for teachers to adhere to the precepts of classical test design theory. Classroom assessments and tests are usually aligned with the informal curriculum and reflect what has been taught or should be learned in the classroom. Teachers often feel that district wide assessments they have helped to design are more reflective of the taught or informal curriculum and thus are more valid than those that are aligned only with national or state curricular objectives that are more reflective of the written or formal curriculum. In other words, assessments and tests may be designed to measure competence to objectives of a formal state or subject area curriculum such as NCTM or NCTE, or to the local curriculum, or to all three. This paper examines the results of a kindergarten assessment designed at the district level by a team of teachers, consultants, and curriculum specialists for an urban school system with a majority of ethnic minority students almost half of whom were eligible for free and reduced meals.

Although there are arguments that constructing such a test is a difficult task, both classical test theory and the more recent development of item response theory are posited on several underlying concepts based on decades of research on principles of good test design, one of which is the concept of unidimensionality (Camilli & Shepard, 1994; Hambleton, Swaminathan & Rogers, 1991; Humphreys, 1986). However it is often the case that longer tests may measure multiple related skills or knowledge areas and thus violate the assumptions of good test design, as tests that measure more than one trait may also yield low reliability coefficients of internal consistency or exhibit other characteristics of multidimensionality. In some cases it is possible to construct a longer assessment of several shorter and related unidimensional tests variously called testlets, parcels, or item blocks. Although it is not always advisable to use the testlet approach, depending upon the purpose of the assessment, it enables counselors and test developers to assess multiple and related areas of student knowledge in a relatively efficient way. Testlets have been variously defined, but are generally considered to be item blocks of four or five items that have been identified through psychometric and statistical procedures as a unit of analysis to measure a single ability, cognitive domain, or latent trait that can be predicted by the student's observed score or performance on the testlet (Berger & Veerkamp, 1994; DeVellis, 1991; Lee, 1998; Lee, Dunbar, & Frisbie, 1999; Thompson & Melancon, 1996; Wainer & Keily, 1987; West, Denton, & Germino-Hausken,. 2000).

The purpose of this paper is to examine the items of a kindergarten assessment to determine if the extraction of component testlets provides additional information beyond that of examination of individual items that would be helpful to teachers, counselors, administrators, and parents in making decisions about children's developmental stage and in early identification of special needs or exceptionality.

Content Validity

A twenty item Kindergarten Diagnostic performance assessment was designed by early childhood curriculum experts in an urban school district to include items that, according to accepted standards of early childhood education, would measure developmentally appropriate academic math and language skills included as part of the curriculum for most preschool students but that were also reflective of the district's curriculum framework (Engel, 1991; Kamii, Manning, & Manning, 1991; Shepard, Kagan, & Wurtz, 1998; Snow, Burns, & Griffin, 1998; West et al., 2000; Wortham, 1995). Even though all teachers assess student progress informally on a daily basis, studies suggest that without the structure provided by a formal assessment, teachers have very low false positive rates but high false negative rates for predicting children at risk for later reading failure, missing perhaps 59% of students who later were identified as disabled readers, thus providing further empirical evidence of the importance of standardized assessments for early literacy (Flynn & Rahbar, 1998). While the majority of items on the diagnostic assessment measured cognitive academic language skills (phonemic awareness, oral fluency, and letter recognition), several items required students to demonstrate counting blocks, drawing geometric figures, grouping numbers of items, and telling number stories. The $r_{20}$ scales for subgroup populations on the twenty item assessment ranged from .87 to .91, generally considered an indication of acceptable internal consistency or reliability across tested populations on the test administration examined in this study, although there are some pitfalls associated with the use of this statistic as shall be discussed later in this paper. However, because the items of the assessment appeared to measure more than one construct, it was of interest for the purposes of this study to identify and analyze factors that represented clusters or domains of skill areas (testlets). Such an analysis would not only help in the identification and remediation of children with developmental or cognitive delays in specific areas of learning, or who might be later identified as gifted, it would also assist the test developers to assure that the test measured what it was designed to measure, and that inferences based on scores would be valid for the purpose for which they would be put to use (Rudner, L. M., 1994; Scott, Deuel, Jean-François, & Urbano, 1996; Scott, Deuel, Urbano, Fletcher, & Torres, 1998; Snyder, 1993; Thompson & Melancon, 1996; Wenner, 1995).

Although this assessment was subsequently replaced by a longer state wide assessment measuring only phonetic and verbal constructs, the ability to perform the items on the Kindergarten Diagnostic Assessment had been determined by the school system to be essential to the demonstration of cognitive academic language proficiency (CALP) for kindergarten activities and reflective of the district's curricular framework. The language related items were consistent with items found on many early childhood assessments, such as ability to retell a story, associating letters and sounds, word recognition, book handling, following print on a page, making predictions about stories, and knowing letter sounds (Bezruczko & Zell, 1993; Engel, 1991; West, Denton, & Germino-Hausken, 2000). Kaminski and Good's study (1996) examined three dynamic indicators of basic early literacy skills (DIBELS), (a) phonemic segmentation fluency, (b) letter naming fluency, and (c) picture naming fluency for a group of kindergarten and first grade students and found that DIBELS were predictive of future reading performance although it was more difficult to determine a causal relationship between the DIBELS and later

success in reading. Various studies have identified phonological awareness as a more significant predictor of early literacy than oral language although oral language ability was identified as a predictor for other developmental areas in children (Bryant, MacLean, & Bradley, 1990; Dickinson & Snow, 1987; Nurss, 1987; Nurss & McGauvran, 1986; Pflaum, 1974; Share, Jorm, MacLean & Matthews, 1984; Snyder & Downey, 1991; Speece, Roth, Cooper, & de la Paz, 1999). Researchers have also sought to establish connections between number recognition, early literacy, and reading skills and thus the inclusion of test items that required children to identify figures and numbers might be expected to yield additional information of interest to teachers and counselors in predicting readiness for kindergarten (Dermott, Pinzari, Dodd, & McIntire, 1980; Kirby & Parrila, 1999).

Extant literature and test reviews suggested that factor analysis of items on the 20 item assessment might be expected to result in three to five item clusters or testlets measuring spacial skills or figure recognition, measurement of numbers of items (counting), phonics and word sense, and verbal ability in speaking, retelling, and describing pictures. One item on the assessment required that students be able to identify as well as draw a geometric figure, thus a correct response to the item represented a measure not only of cognitive ability but also of fine motor skills. The present analysis posited that the extraction and identification of factor loadings of the items on the assessment would yield helpful information about item clusters and also about how children learn (Knight, 2000).

## Testlet Development

Testlets, as the name implies, have been defined variously as small tests that are contextually consistent, or groups of related items within a single content area (Lee, 1998; Thompson & Melancon, 1996; Wainer & Kiely, 1987). When test items violate the assumption of independency, the grouping of dependent items into independent testlets is an accepted approach to avoid or diminish the influence of dependency between items (Lee, 1998; Lee, Dunbar, & Frisbie, 1999; Thissen, Steinberg & Mooney, 1989). The testlet was initially the concept of Wainer and Keily (1987) who conceived its use in computer adaptive testing and who defined it as "a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow" (p. 190). In more current usage testlets are not limited to computer testing but generally refer to item pools. Wainer, Kaplan, and Lewis (1991) extended their analyses to determine the optimal length of test, testlet, and item pool for linear and hierarchical testlets. Their study examined testlets of 4, 5, 6, 7, and 8 items and item pools of 15, 30, 50, and 100 items. They found that testlets of 4 to 8 items from an item pool of as few as 15 good items yielded satisfactory results. If tests are designed such that they are composed of interdependent items or testlets, then reliability coefficients computed as if all items were independent may yield a different (and higher) reliability coefficient than if reliability were computed after controlling for within testlet dependencies (DeVellis, 1991; Lee, 1998; Wainer, Sireci, & Thissen, 1991). Lee, Brennan, and Frisbie (2000) further expanded and simplified the definition of testlet to simply be defined for their purposes as "a measurement unit" (p. 10) that could include single items, such as a group of questions about one figure or graph on a test. Their new definition would allow for testlets to be either locally dependent or locally

independent, a departure from the most commonly accepted practice of classical test design theory. They further recommended that the testlet should be the unit of measure in test design unless it could be shown that the testlet unit would be a detriment or not beneficial to the analysis of the psychometrics of the test (Lee et al., 2000). Therefore, an examination of the dimensionality of the kindergarten assessment using the testlet approach was of interest to determine the extent to which the assessment was either unidimensional or multidimensional and if the identification of multiple factors was demonstrably beneficial in developing a kindergarten readiness profile for children.

## Methodology and Results

The instrument was a 20 item assessment administered to 1137 prekindergarten students in an urban school district composed of 47% Black, 27% Hispanic, 19% White, and 7% other ethnic groups, 127 of whom were identified as Limited English Proficient. Fifty-five percent of incoming kindergarten students were entitled to free or reduced price meals. The prekindergarten students ranged from 61 to 102 months of age, with a mean age of 67 months. On the average, older students up to 7 years of age did better on the assessment than younger students and there was no significant correlation between age and ethnicity. Although beyond the scope of this paper there was also a significant school effect after controlling for other variables and LEP students who were eligible for free or reduced price meals performed better on the assessment than LEP students who were not eligible for free or reduced price meals. For the purposes of this analysis only scores of LEP or ESL students were excluded. Raw test scores for the total test for the population of students were approximately normally distributed ($\underline{M}$ = 9.4, SE = .15, $\underline{SD}$ = 4.7). $KR_{20}$ for the 20 item total test yielded a reliability of .87 for 961 non-LEP kindergarten students with complete data on all the variables; however, there was a high degree of dependency between some items, with a maximum inter-item correlation of .73. Given the indication that the test indeed exhibited some characteristics of multidimensionality as well as local dependency, a factor analysis and identification of testlets or item blocks was deemed appropriate, particularly to assure that test reliability was not inflated due to dependency among items and to compare inter-item and inter-testlet dependency and reliability coefficients. The proposed nonrecursive model is provided in Figure 1.

---

Insert Figure 1 about here

---

Because there was a theoretical basis for the assumption that the kindergarten assessment was multidimensional the present analysis essentially sought to identify groups of intercorrelated item testlets, first through preliminary EFA with followup CFA procedure. For the 20 items on the assessment for a population of 961 non-LEP kindergarten students, the result of the Kaiser-Meyer-Olkin statistic of .895 was meritorious and indicated a strong degree of correlation between the observed and partial correlation coefficients with the Bartlett Test of Sphericity rejected with a $\chi^2$ (190) = 6189.71, $p < .001$. The magnitude of the KMO test and very small significance level for the large value of the $\chi^2$ was indicative of a multivariate normal population, correlations between the variables related to other variables and that factor analysis could proceed

(Norusis, 1992; Tabachnic & Fidell, 1989). A Principle Component Analysis of the 20 item assessment using Varimax rotation with Kaiser Normalization converged in eight iterations and resulted in a four factor solution (Table 1), with Factor One and Two producing six item testlets, and Factor Three a four item testlet. One item was omitted from the subsequent analyses as it did not appear to load definitively on any of the factors. Three items loaded in Factor Four, although only one item was identified as loading above a value of .5. There are caveats in the literature related to over extraction of components and the difficulties that result when over extraction leads to failure of models to converge due to large numbers of components (Lawrence & McLean, 2000). Lee (1998) also noted that the Kaiser criterion has been criticized for overidentifying dimensions and cites Stout's contention that the consideration of unidimensionality may be satisfied by the identification of one dominant component or factor and perhaps several minor factors that he termed essential unidimensionality. Thus a test may be considered unidimensional if it contains a dominant component with additional information provided by the identification of minor factors. Lee et al. (2000) further assert that testlet items may be either locally dependent or locally independent, depending upon the knowledge area or construct to be measured. The results of the PCA indicated that the first four factor components accounted for 53.65% of variance, with Factor One explaining approximately 30% of variance, Factor Two 9.6% of variance, Factor Three 6.92% of variance, and Factor Four 4.60% of variance. Thus, the alternative hypothesis that the 20 item assessment was unidimensional was rejected and a CFA of the four testlets identified by the PCA was conducted with the following results.

---

Insert Table 1 about here

---

Covariance coefficient matrices of examinee scores on the four testlets identified by PCA were analyzed using LISREL Maximum Likelihood estimation to test model fit. The results of model fitting for the four factor model are provided in Figures 2 through 4. A preliminary nonrecursive four factor model (Figure 2) converged in four iterations but was poorly fitting with an RMSEA of .11 and a p-value for $\chi_2$ of <.001 and was rejected. Although the model in Figure 3 converged with the addition of a covariance between Factor One and Factor Four and yielded a small $\chi_2$, the relatively small p-value for $\chi_2$ of .012 and the RMSEA (.074) larger than the recommended .05 led to rejection of this model (Table 2).

---

Insert Figures 2 - 5 about here

---

Figure 4 shows the model with an additional covariance from Factor One to Factor Three, resulting in a fully saturated, perfectly fitting recursive model. All four factors could be fitted to the model by adding covariances between factors although this somewhat defeated the purpose of identifying testlets. The model was subsequently tested without Factor Four; this was deemed a reasonable procedure for several reasons: to minimize dependencies between factors and also because Factor Four contained only three items while recommendations suggest that a testlet should consist of at least four items (Wainer, Kaplan, et al., 1991) or, alternatively, one reading passage or figure accompanied by related questions. The three factor model (Figure 5) resulted in

a nonrecursive fully saturated model, with Factor One $\underline{R}^2$ = .40, Factor Two $\underline{R}^2$ = .45, and for Factor Three $\underline{R}^2$ = .47.

The final model was purposely relatively simple, with few parameters. However, there is a limitation connected with models with few parameters, as chi-square becomes smaller when there are more parameters. As a result, models with few parameters may fail to meet significance criteria for chi-square while models with over extracted parameters (to yield smaller chi-squares) may fail to converge the majority of the time (Lawrence & McLean, 2000). This difficulty calls attention to the advisability of examining additional fit indices when fitting the model, as well as to parsimony in extraction of components and addition of parameters. Table 2 provides additional measures of fit for several of the tested models for comparison purposes. Several of these goodness-of-fit measures consider the number of parameters or parsimony of the model. Unlike chi-square, the AIC favors models with fewer parameters. The ECVI measures the difference between the fitted and expected covariance matrix given another sample of the same size. These fit measures rank order models and the model with the smallest value is selected.

---

Insert Table 2 about here

---

With kindergarten readiness as the latent variable, the model yielded the following equations for the three factors:

| | |
|---|---|
| FAC1 =1.23*Readiness, | Error variance = 2.27 |
| (.071) | (.15) |
| 17.18 | 15.15 |
| FAC2 = 1.17*Readiness, | Error variance = 1.69 |
| (.065) | (.13) |
| 17.97 | 13.47 |
| FAC3 = 0.93*Readiness, | Error variance = 0.99 |
| (.051) | (.077) |
| 18.23 | 12.86 |

---

Insert Table 3 about here

---

Table 4 provides inter-testlet correlations, demonstrating lower inter-testlet correlations (maximum .55) than were the inter-item correlations (maximum .73) when scaled as 20 individual items. The standardized item alpha for the four testlet assessment was .77, somewhat less than the .87 of the 20 item scale, consistent with the findings of DeVellis (1991), Lee (1998, 1999), and Wainer, Sireci et al. (1991) The t-test for unidimensionality failed to reject the hypothesis at the .05 level of significance that the testlets measured one dominant factor for each of the three testlets that were used in the structural equation model (Table 5).

---

Insert Tables 4 and 5 about here

---

## Discussion

Consistent with the literature on interpretation of factors and components, while the identification of the dominant factor of two of the testlets that were developed in the study were readily apparent, the dominant factor of the other two was somewhat problematic. The Factor One testlet consisted of items that measured identification of letters, shapes, and consonant sounds. That there is a connection between ability to recognize the shapes of upper and lower case letters of the alphabet and ability to recognize common geometric shapes is somewhat intuitive, although the phonemic sound production items did not appear so readily to belong in the same testlet as letter and figure shapes. However, if one assumes that much of the underlying construct measured is related to instruction, with letters and their sounds taught concurrently, this dimension becomes more understandable. This information could help provide teachers with means to identify children who might have specific types of developmental delays. The items of Factor Two, a six item testlet, were primarily items that measured verbal fluency, such as ability to explain a concept, to retell a story or tell a story about a picture. Once again, the use of a testlet to measure this construct rather than individual test items would have the potential to provide teachers with a scale to measure a child's fluency in the use of language. The items of Factor Three, a four item testlet, were all measures of counting and grouping numbers of blocks, both individually and in sets. The three items of Factor Four (book handling) were less apparently consistent in content than either Factor Two or Factor Three, with two items that were measures of see and say skills and one item that consisted of multiple steps: children were asked to hold a book correctly, identify the title page, and follow print on the page from left to right. However, once again the underlying construct could be informative for identification of children who were or who were not accustomed to handling books.

In conclusion, although reliability coefficients indicated that scale reliability was lower for the four item testlets than for the same test scaled as 20 individual items, the reliability coefficient was higher than expected given a scale of only four items. This result was consistent with the findings of Lee et al. (2000). However, the additional utility provided by being able to report student readiness along a continuum of each of four scales as well as on the total 20 item assessment would more than offset the small reduction in reliability. Because the items on the 20 item assessment were dichotomous, scored as either correct or incorrect, the information provided by an incorrect response by a child on an item was very limited as there was no mechanism to determine "how wrong" an incorrect response was, although a correct response indicated complete mastery of the criterion. Each item required the child to complete several tasks, such as name and draw a circle, square, triangle, and rectangle. Thus, a child might know the circle or square or be able to identify several letters of the alphabet, but might not meet the criteria for getting the item correct. The use of testlets and testlet scores in assessment provides additional and helpful information about children's developmental progress and readiness for kindergarten and might also aid in early identification and remediation of undiagnosed learning difficulties (Wenner, 1995).

Future research might examine differences on the identified testlets for different populations of kindergarten and first grade students to determine if testlet scores have predictive validity for reading readiness and success in kindergarten as well as to determine if groups of students do better on some testlets than on others.

Bibliography

Berger, M. P. F & Veerkamp, W. J. (1994). A review of selection methods for optimal test design. Research Report 94-4. (ERIC Document Reproduction Service No. ED 389 755)

Bryant, P., MacLean, M., & Bradley, L. (1990). Rhyme, language, and children's reading. Applied Psycholinguistics, 9, 367-383.

Camilli, G., & Shepard, L. (1994). MMSS: Methods for identifying biased test items. Thousand Oaks, CA: Sage Publications, Inc.

Dermott, R. A., Pinzari, S., Dodd, E. R., & McIntire, W. (1980). Number knowledge for prediction of specific reading skills and achievement in first graders. Reading Improvement, 17, 86-93.

DeVellis, R. F. (1991). Scale development: Theory and applications. Applied social research methods series (vol. 26). Newbury Park: Sage Publications.

Dickinson, D. E., & Snow, C. E. (1987). Interrelationships among the prereading and oral language skills in kindergarten children from two social classes. Early Childhood Research Quarterly, 2, 1-25.

Engel, B. S. (1991). An approach to assessment in early literacy. In C. Kamii, M. Manning, & G. Manning (Eds.). Early literacy: A constructivist foundation for whole language. NEA Early Childhood Education Series. Washington, DC: National Education Association of the United States.

Flynn, J. M., & Rahbar, M. H. (1998). Improving teacher prediction of children at risk for reading failure. Psychology in the Schools, 35(2), 163 - 172.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. Journal of Applied Psychology, 71, 327-333.

Kamii, C., Manning, M., & Manning, G. (1991). Early literacy: A constructivist foundation for whole language. NEA Early Childhood Education Series. Washington, DC: National Education Association of the United States.

Kaminski, R. A., & Good, R. H. III. (1996). Toward a technology for assessing basic early literacy skills. School Psychology Review, 25(2), 215 - 227.

Kirby, J. R., & Parrila, R. K. (1999). Theory-based prediction of early reading. Alberta Journal of Educational Research, 45(4), 428-447.

Knight, J. (2000). Toward reflective judgment in Exploratory Factor Analysis decisions: Determining the extraction method and number of factors to retain. Paper presented at the Annual Meeting of the Mid-South Educational Research Association. (Bowling Green, KY, November 15-17, 2000).

Lawrence, F. R., & McLean, J. E. (2000). Effects of over extraction in finite mixture growth models. Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Bowling Green, KY, November 15-17, 2000).

Lee, G. (1999). Conditional standard errors of measurement for tests composed of testlets. Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 19-23, 1999). (ERIC Document Reproduction Service No. ED 439 143)

Lee, G. (1998). Estimating conditional standard errors of measurement for tests composed of testlets. Paper presented at the Annual Meeting of the Iowa Educational Research and Evaluation Association.(Ames, IA, December 4, 1998).(ERIC Document Reproduction Service No. ED 439 114)

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analysis. Educational Measurement: Issues and Practice 19(4), 9-15.

Lee, G., Dunbar, S. B., & Frisbie, D. A. (1999). Measurement models for a testlet-based test. Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 19-23, 1999). (ERIC Document Reproduction Service No. ED 439 145)

Norusis, M. (1992). SPSS/PC+ professional statistics: Version 5. Chicago: SPSS Inc.

Nurss, J. (1987). Readiness for kindergarten. ERIC Digest. (ERIC Document Reproduction Service No. ED 291 514)

Nurss, J. R., & McGauvran, M. E. (1986). Metropolitan Readiness Tests. San Antonio: The Psychological Corporation.

Pflaum, S. W. (1974). The development of language and reading in the young child. Columbus, OH: Charles E. Merrill.

Rudner, L. M. (1994). Questions to ask when evaluating tests. Practical Assessment, Research, & Evaluation, 4(2). Available online: http://ericae.net/pare/getvn.asp?v=4&n=2

Scott, M. S., Deuel, L. S., Jean-François, B., & Urbano, R. C. (1996). Identifying cognitively gifted ethnic minority children. Gifted Children Quarterly, 40(3), 147-53.

Scott, M. S., Deuel, L. S., Urbano, R. C., Fletcher, K. L., & Torres, C. (1998). Evaluating the initial version of a new cognitive screening test. Education and Training in Mental Retardation and Developmental Disabilities, 33(3), 280-289.

Share, D. L., Jorm, A. F., MacLean, R., & Matthews, R. (1984). Sources of individual differences in reading achievement. Journal of Educational Psychology, 76, 1309-1324.

Shepard, L., Kagan, S. L, & Wurtz, E.(Eds.).(1998). Principles and recommendations for early childhood assessment. National Education Goals Panel. Washington, D. C.: U. S. Government Printing Office.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28(3), 237-247.

Snow, C., Burns, S, & Griffin, P. (1998). Preventing reading difficulties in young children. Washington, D.C.: National Academy Press.

Snyder, L. S., & Downey, D. M. (1991). The language-reading relationship in normal and reading-disabled children. Journal of Speech and Hearing Research, 34, 129-140.

Snyder, S. (1993). Program evaluation: DEC Recommended Practices.(ERIC Document Reproduction Service No. ED 370 267)

Speece, D. L., Roth, F. P., Cooper, D. H., & De La Paz, S. (1999). The relevance of oral language skills to early literacy: A multivariate analysis. Applied Linguistics, 20, 167-190.

Tabachnic, B., & Fidell, L. S. (1989). Using multivariate statistics (2nd ed.). New York: Harper Collins.

Thompson, B., & Melancon, J. G. (1996). Using item "Testlets"/"Parcels" in Confirmatory Factor Analysis: An example using the PPSDQ-78. Paper presented at the Annual Meeting of the Mid-South Educational Research Association. (Tuscaloosa, AL, November 1996). (ERIC Document Reproduction Service No. ED 404 349)

Wainer, H., Kaplan, B., & Lewis, C. (1991). Some empirical guidelines for building testlets. Program Statistics Research Technical Report No. 91-14. Princeton: Educational Testing Service.

Wainer, H., & Keily, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185-201.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. Journal of Educational Measurement, 28(3), 197-219.

Wenner, G. (1995). Kindergarten screens as tools for the early identification of children at risk for remediation or grade retention. Psychology in the Schools, 32(4), 249-254.

West, J., Denton, K., & Germino-Hausken, E. (2000). America's kindergartners: Early Childhood Longitudinal Study - Kindergarten class of 1998-1999. Report No. NCES 2000-070. National Center for Education Statistics. Washington, DC: U.S. Department of Education Office of Educational Research and Improvement.

Wortham, S. C. (1995). Measurement and evaluation in early childhood education (2nd ed.). Englewood Cliffs, N. J.: Merrill.

Table 1

Factor Loadings and Communalities for Confirmatory Factor Model of Kindergarten Readiness (n = 962)

| Measure and variable | Rescaled factor loading | Communalities |
|---|---|---|
| FAC1 (Letter sound and figure recognition) | | |
| Uppercase letters | .841 | .761 |
| Lowercase letters | .811 | .753 |
| Letters represent sounds | .701 | .590 |
| Beginning consonant sounds | .686 | .559 |
| Writes beginning consonants | .612 | .473 |
| Name and draw geometric figures | .545 | .462 |
| FAC2 (Verbal fluency) | | |
| Retells | .748 | .619 |
| Makes predictions about stories | .717 | .543 |
| Recognizes patterns in text | .618 | .451 |
| Demonstrates knowledge gained from pictures | .605 | .407 |
| Distinguishes real, make-believe | .531 | .394 |
| Verbalizes ideas | .481 | .325 |
| FAC3 (Math counting) | | |
| Associates number of objects with a number | .756 | .652 |
| Uses nonstandard units of measure | .755 | .640 |
| Tells number story | .687 | .525 |
| Tells number in a set of items | .655 | .496 |
| FAC4 (Awareness of printed materials) | | |
| Book handling, follows print on page | .763 | .719 |

Note. Communality ( $h^2$ statistic) is the squared multiple correlation with its factors.

Table 2

Fit Indices for the Tested Models

| | $\chi^2$ | df | GFI | AGFI | RMSEA | NFI | CFI | PNFI | PGFI |
|---|---|---|---|---|---|---|---|---|---|
| Nonrecursive 4 factor | 24.27 | 2 | .99 | .94 | .11 | .97 | .98 | .32 | .20 |
| Partially recursive 4 factor | 6.31 | 1 | 1.00 | .97 | .07 | .99 | .99 | .17 | .10 |
| Saturated 4 factor | 0 | 0 | 1.00 | – | – | – | – | – | -- |
| Saturated 3 factor | 0 | .0 | 1.00 | – | – | – | – | – | -- |

Note. GFI, AGFI, RMSEA, NFI, CFI, PNFI, and PGFI refer to goodness-of-fit index, adjusted goodness-of-fit index, root mean square error of approximation, normed fit index, comparative fit index, parsimony normed fit index, and parsimony goodness-of-fit index.

Table 3

Standardized Parameter Estimates for the Measurement Model

| | Kindergarten Readiness | $R^2$ |
|---|---|---|
| Factor One: Letters, Shapes, Sounds | .66 | .40 |
| Factor Two: Verbal Fluency | .74 | .45 |
| Factor Three: Counting and Sets | .66 | .47 |

Table 4

Correlations of the Kindergarten Assessment Testlets

| | Factors | | | |
|---|---|---|---|---|
| Factors | Letters, Sounds, Shapes | Verbal Fluency | Counting Sets | Book Handling |
| Letters, Sounds, Shapes | 1.00 | | | |
| Verbal Fluency | .450** | 1.00 | | |
| Counting and Sets | .456** | .491** | 1.00 | |
| Book Handling | .553** | .492** | .419** | 1.00 |

Note. ** $p < .01$

Table 5

T-Test for Unidimensionality of Testlets

| | | | T Statistics | |
|---|---|---|---|---|
| Testlet | N items | N examinees | T-value | Probability |
| Figures, Letters, Sounds | 6 | 961 | 17.18 | .071 |
| Verbal Fluency | 6 | 961 | 17.97 | .065 |
| Counting and Sets | 4 | 961 | 18.23 | .051 |

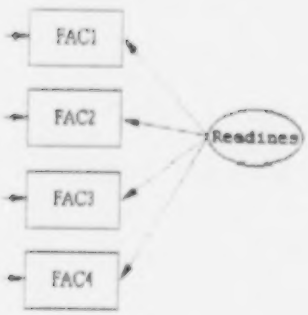



Figure 1. Proposed four factor nonrecursive model for factors predicting latent variable kindergarten readiness.

BEST COPY AVAILABLE

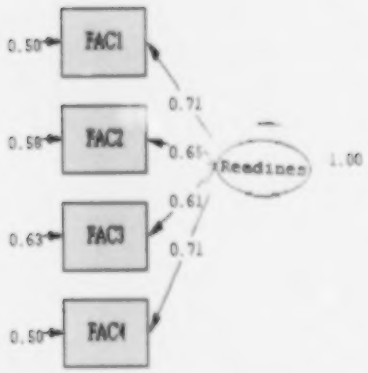
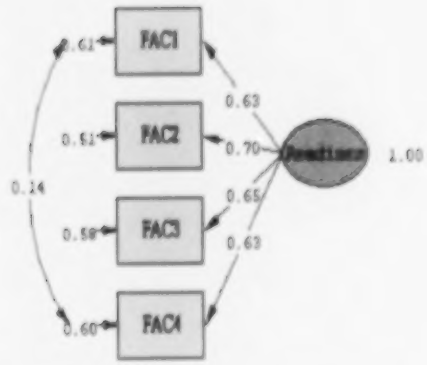



Figure 2. Preliminary nonrecursive standardized estimates for a four factor model demonstrating poor fit indices.

BEST COPY AVAILABLE




Chi-Square=6.28, df=1, P-value=0.01218, RMSEA=0.074

Figure 3. Standardized estimates for recursive four factor model with poor fit indices.

BEST COPY AVAILABLE

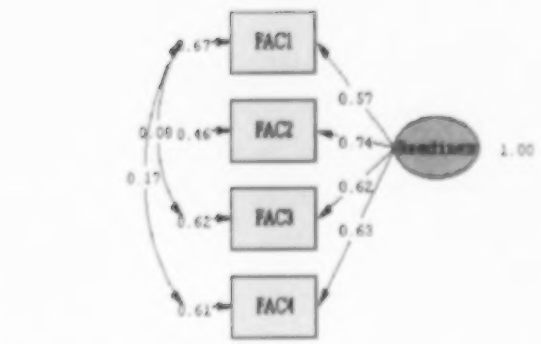



Figure 4. Standardized solution for recursive four factor saturated model.

BEST COPY AVAILABLE

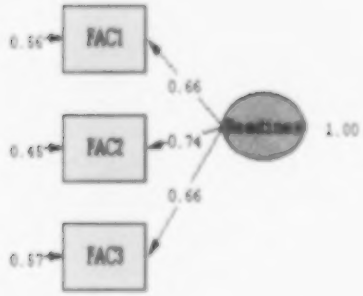



Figure 5. Nonrecursive three factor saturated independence model with standardized solution for predictors of latent variable kindergarten readiness.

BEST COPY AVAILABLE